Atomic Energy Control Board
P.O. Box 1046
Ottawa, Canada
K1P 5S9

Commission de contrôle de l'énergie atomique
C.P. 1046
Ottawa, Canada
K1P 5S9

INFO-0417-1

92-05774/

OCCUPATIONAL RADIATION EXPOSURE AND MORTALITY STUDY: I INTERNAL LINKAGE OF THE CANADIAN NATIONAL DOSE REGISTRY

(AECB Project No. 7.101.1)

by

E. Coppock, D. Dobson
and
M. Fair

Statistics Canada

A research report prepared for
the Atomic Energy Control Board
Ottawa, Canada

June 1992

Canada

Research report

OCCUPATIONAL RADIATION EXPOSURE AND MORTALITY STUDY: I. INTERNAL LINKAGE OF THE CANADIAN NATIONAL DOSE REGISTRY

A report prepared by Elizabeth Coppock, Danna Dobson and Martha Fair, Occupational and Environmental Health Research Section, Canadian Centre for Health Information, Statistics Canada, under contract to the Atomic Energy Control Board.

ABSTRACT

An epidemiological cohort study of some 300,00 Canadians enroled in the National Dose Registry (NDR) is being undertaken to determine if there is excess cancer or other causes of mortality among those workers who are occupationally exposed to low levels of ionizing radiation. The results of this study may provide better understanding of the dose-response relationship for low doses of ionizing radiation and aid in the verification of risk estimates for radiation-induced cancer mortality. The Department of National Health and Welfare (DNHW) is responsible for the Registry; this study is being carried out by the Bureau of Radiation and Medical Devices (BRMD) with financial assistance and co-operation of various agencies including Statistics Canada and the Atomic Energy Control Board.

This report is prepared, as part of the contractual agreement with the Atomic Energy Control Board, to give details of the work carried out at Statistics Canada and focuses on the first major task — assembling the study cohort.

Five files of data totalling 439,459 records (not individuals) were provided to Statistics Canada by Bureau of Radiation and Medical Devices. To expedite the computerized record linkage of these files to the Canadian Mortality Data Base (CMDB), it was necessary to improve the quality of information on the National Dose Registry by maximizing the availability of information for fields vital to the linkage, and to remove duplicate records relating to the same individual. To do this, fragmented personal histories due to job changes and duplicate entries due to name changes of individuals in the National Dose Registry were brought together under a unique numeric identifier. The cohort of individuals resulting from this "internal linkage" of the National Dose Registry records proceeds to the mortality linkage with the Canadian Mortality Data Base (Report II).

The Master Identification File (MIF) of 439,459 records underwent an evaluation for quality and quantity of personal identifying data. Subsequent to this initial editing step, 18,057 historic records from Atomic Energy of Canada Limited were added to the Master Identification File. Because a high percent of the records in the Master Identification File had missing information, it would, in many cases, be difficult to identify persons unambiguously during the linkage with the Canadian Mortality Data Base. To improve the quality of the linkage, it was decided to use the Social Insurance Number (SIN) Index to aid in individual identification.

After the internal linkage, which involved several stages, and the final updates were completed, there were 266,559 individuals in the cohort for the mortality linkage. Of these, 143,396 were males and 123,814 were females. There were 651 individuals without a sex code — these were added to both males and females for the mortality linkage. There were 64,007 individuals who did not have the minimum identifiers necessary to participate in the mortality linkage at this time. Further investigation of this file will be carried out.

RÉSUMÉ

Le présent rapport porte sur une étude épidémiologique d'une cohorte d'environ 300 000 Canadiens inscrits dans le Fichier dosimétrique national. L'étude est entreprise afin de déterminer s'il y a un nombre plus élevé de cas de cancer mortel ou de d'autres causes de décès chez les travailleurs qui sont exposés à de faibles niveaux de rayonnement ionisant dans leur milieu de travail. Les résultats de l'étude permettront peut-être de mieux comprendre la relation dose-réponse pour les faibles doses de rayonnement ionisant et d'aider à vérifier les estimations du risque de cancer mortel attribuable aux rayonnements. Le ministère fédéral de la Santé et du Bien-être social est responsable du Fichier dosimétrique national. La présente étude est effectuée par le Bureau de la radioprotection et des instruments médicaux, avec l'aide financière et la collaboration de divers organismes dont Statistique Canada et la Commission de contrôle de l'énergie atomique.

Le présent rapport s'inscrit dans le cadre d'un contrat conclu avec la Commission de contrôle de l'énergie atomique en vue de fournir le détail des travaux menés à Statistique Canada et porte sur la première tâche importante, soit l'établissement de la cohorte à l'étude.

Le Bureau de la radioprotection et des instruments médicaux a fourni à Statistique Canada cinq fichiers de données contenant un total de 439 459 fiches (et non des personnes). Pour accélérer le couplage informatisé de ces fiches à la Base canadienne de données sur la mortalité, il a d'abord fallu améliorer la qualité de l'information dans le Fichier dosimétrique national en maximisant la disponibilité des renseignements sur les zones essentielles au couplage, puis supprimer les doublets pour une même personne. À cette fin, on a réuni, sous un même identificateur, les antécédents personnels fragmentés en raison des changements d'emploi et les doublets attribuables aux changements de nom des personnes dans le Fichier dosimétrique national. La cohorte de personnes issue de ce «couplage interne» des enregistrements du Fichier dosimétrique national passera ensuite au couplage des données sur la mortalité (rapport II).

Les données d'identification du Fichier permanent d'identification des 439 459 fiches ont fait l'objet d'une évaluation qualitative et quantitative. À la suite de cette première étape de contrôle, 18 057 fiches historiques d'Énergie atomique du

Canada limitée ont été ajoutées au Fichier permanent d'identification. Comme il manquait des renseignements pour un pourcentage élevé d'enregistrements du Fichier permanent d'identification, il aurait été difficile, dans de nombreux cas, d'identifier clairement les personnes durant le couplage avec la Base canadienne de données sur la mortalité. Pour améliorer la qualité du couplage et faciliter l'identification des personnes, on a donc décidé d'utiliser le fichier des numéros d'assurance sociale.

Après le couplage interne, qui a nécessité plusieurs étapes, et les mises à jour finales, la cohorte en vue du couplage des données sur la mortalité comptait 266 559 personnes, dont 143 396 hommes et 123 814 femmes. On a également constaté qu'aucun code de sexe n'avait été attribué dans le cas de 651 personnes; ces personnes ont été ajoutées aux hommes comme aux femmes pour le couplage des données sur la mortalité. Un total de 64 007 personnes n'avaient pas d'identificateur miniml nécessaire au couplage des données sur la mortalité. Une recherche plus approfondie concernant le contenu du fichier sera entreprise.

DISCLAIMER

The Atomic Energy Control Board is not responsible for the accuracy of the statements made or opinions expressed in this publication and neither the Board nor the authors assume liability with respect to any damage or loss incurred as a result of the use made of the information contained in this publication.

TABLE OF CONTENTS

ABSTRACT iii

GLOSSARY ix

A. BACKGROUND 1

1. Introduction 1
2. Purpose 1
3. Composition of the National Dose Registry 2
4. Major Stages of the Record Linkage Phase 4

B. METHODS 4

1. Stage I: Evaluating the Master Identification File 4
   1.1 Contents 4
   1.2 Editing of Identifiers on the Master Identification File 6
   1.3 Availability of Identifiers 6

2. Stage II: Correcting the Master Identification File 6

3. Stage III: Internal Linkage of the Master Identification File 8
   3.1 Linkage By Numeric Identification Numbers . 8
   3.2 Linkage By Numeric Identification Using Non-Numerical Identifiers 11
   3.3 Adding Atomic Energy of Canada Limited Records to the Master Identification File 13
   3.4 Collapsing the New Master Identification File Records 14
   3.5 Master Identification File - Social Insurance Number Index File Linkage 15
   3.6 Linkage of the New Master Identification File by NYSIIS Code 19

4. Stage IV: Merging of Master Identification File with Dose History Data 22

5. Future Work Stages V and VI 23

REFERENCES 24

ACKNOWLEDGEMENTS 26

APPENDIX A

CODES FOR EMPLOYMENT-RELATED VARIABLES SPECIFIC TO NATIONAL DOSE REGISTRY A1

APPENDIX B

QUALITY OF DATA (1.2) EDIT RULES B1

APPENDIX C

COMPARE RULES: INTERNAL LINKAGE BY NUMERICAL IDENTIFICATION USING NON-NUMERICAL IDENTIFIERS C1

APPENDIX D

COMPARE RULES FOR MASTER IDENTIFICATION FILE - SOCIAL INSURANCE NUMBER INDEX LINKAGE D1

APPENDIX E

COMPARE RULES FOR NYSIIS LINKAGE E1

GLOSSARY

The following are acronyms and a glossary of terms frequently used in this report.

AECB - Atomic Energy Control Board, the federal regulatory agency.

BRMD - Bureau of Radiation and Medical Devices, the group at the Department of National Health and Welfare responsible for the operation of the National Dose Registry.

CMDB - Canadian Mortality Data Base, a computer file of all deaths occurring in Canada dating back to 1950 with underlying cause of death.

Combseq - a sequential number assigned to an alphabetically sorted list of individuals who had dose records from 1951 to 1963 and who had no other numeric identifiers.

DNHW - Department of National Health and Welfare.

Dupflag - a indicator for a duplicate record.

GIRLS - Generalized Iterative Record Linkage System, a software system used to provide probabilistic matching of individual records.

Hybrid - a restructured record on the Master Identification File which is composed of information from the original National Dose Registry record and the Social Insurance Index file.

Internal linkage - a computer assisted process by which records within a single file are compared. The process is used to remove duplication and maximize the amount of information available for an individual.

LDHS - Lifetime Dose History System, the data base which currently contains all identifying information and annual summary records for each individual in the National Dose Registry.

Linkage - a computer assisted process by which records are compared to determine whether the records refer to a single individual.

MIF - Master Identification File, the original National Dose Registry records containng the basic identifier information given to Statistics Canada for this study.

Namecode - a number assigned to individuals in the National Dose Registry between 1964 and 1972.

NDR - National Dose Registry, a centralized Canadian computer file of occupational radiation exposure records.

NYSIIS - New York State Intelligence Information System, a phonetic coding system designed to bring together alternate spellings of the same surname.

Pocket ID - a field used to sort or subdivide records for comparisons where each record within a pocket is compared to each other.

RPB No. - Radiation Protection Bureau number, a number assigned to individuals in the National Dose Registry between

1973 and 1977.

SAS - Statistical Analysis System, a statistical software package used to provide data for quality and quantity analyses.

S.I.N. - Social Insurance Number, a unique number assigned to Canadians.

Threshold - a weight above or below which all linkages are said to be true or false.

Weight - a numeric value which assesses the probability that two records do refer to one individual. Weighting factors are derived from the files themselves by, for example, the frequencies with which different identifiers appear on the files and are based on the greater discriminating power of rare names, etc. In general, higher weighted linkages are more likely to be correct.

A. BACKGROUND

1. Introduction

There is growing concern among both the general public and workers about the health effects of exposure to low doses of ionizing radiation. Of major concern is the risk of developing cancer. The risk of certain cancers increases with increasing radiation exposure; however, this fact is almost totally based on data extracted from studies of individuals exposed to doses higher than occupational doses and doses which are delivered in short periods of time i.e. the Japanese atomic bomb survivors.[1]

There is uncertainty about the actual magnitude of the risk associated with occupational exposures which are much lower and are accumulated slowly over long periods of time (See refs. 1-4 for a discussion of these points). Many epidemiological studies are currently under way in an attempt to resolve this issue[5-7]. Since the effects from very low level chronic exposures such as those received currently by radiation workers are expected to be small, a very large cohort size is necessary for these studies to have sufficient power to detect such small changes[8].

In Canada, large scale epidemiological studies of radiation workers can be undertaken because of the existence of two national facilities - the National Dose Registry and the Canadian Mortality Data Base. The NDR, operated by BRMD within the DNHW, is a comprehensive, centralized, computerized file of virtually all workers in Canada, who have been occupationally exposed to radiation together with their yearly exposure data[9,10]. Some records in this file date from the early 1950's. There are approximately 300,000 individuals in 80 different job categories and covering up to 40 years of exposure data in this file. Canada also has a centralized, computerized national mortality file, the CMDB, dating back to 1950[11]. This file contains the coded underlying cause of death for all persons who have died in Canada since that time. Statistics Canada has the necessary generalized software[12] to conduct studies which would link these two files together in order to determine the fact of death and cause of death in the NDR cohort.

2. Purpose

A study to link these two files was set up in 1984 by BRMD with the assistance of Statistics Canada and the AECB[13]. The purposes of the study are:

1. To compare the mortality of the cohort of workers monitored for ionizing radiation with that of the overall population of Canada.
2. To determine whether or not there is a difference in

cause of death between exposed and non-exposed members of the cohort.

3. To determine if death from specific diseases occurs more frequently in the exposed members of the cohort than in the non-exposed members.
4. To determine if a dose response relationship can be established between the level of exposure and the mortality of the cohort.
5. If no dose response relationship can be established, to determine an upper limit on the residual level of risk associated with radiation exposure.

The methodology used to ascertain the mortality of individuals on the NDR was a computerized record linkage to the CMDB using the Generalized Iterative Record Linkage System (GIRLS)[14,15]. Because of poor identifying information on a portion of records in the NDR, a linkage to the Social Insurance Number (S.I.N.) Index file was carried out to capture more identifying data for these persons so that they could be unambiguously identified[16].

The portion of the work to be done at Statistics Canada is:

1. To prepare the Master Identification File (MIF) for linkage with the CMDB i.e. to bring together all records relating to the same individual.
2. To perform the linkage with the CMDB to determine the fact of death and the underlying cause of death for those individuals in the NDR who have died.
3. To prepare an analysis file for HWC which contains the dose data from the NDR and the death information from the CMDB.
4. To prepare documentation to support the work done at Statistics Canada.

Within the overall plan of the study, this report presents, in detail, the steps that were necessary to improve and verify personal identifying information on the NDR prior to its linkage with the CMDB. Owing to the large size of the file and the varied quality of the data collected over a period of 40 years for purposes other than record linkage, this work took several years and involved many persons at BRMD and Statistics Canada.

3. Composition of the National Dose Registry

At the start of this study in 1984, the National Dose Registry was comprised of basically four files[9,10].

The first file, called the Master Identification File, contained personal identifying information such as surname, given names, sex, year of birth, and assigned identification numbers used to identify the individuals' dose records.

The second file, the Lifetime Annual History File, consisted of a sequential file summarizing each individual's dose history in a series of records, one for each year. In addition to the individual's numerical identifier and dose information, this file contained the province and type of employing organization, identifying serial number of the organization, and the type of work as of the last reported record for the year.

The third file, the Cumulative File, was actually a database which contained cumulative radiation doses for the current, defined by last entry, and previous year as well as the individual's lifetime total dose.

The last file, the Discrete History File, consisted of a microfilm record of each dose entry and dose changes made to an individual's record where applicable.

In 1986, the NDR was restructured to consist of a single database called the 'Lifetime Dose History System' (LDHS) and a Discrete Transaction File. The LDHS contains all the identifying information and annual summary records for each individual who was monitored. The database can be searched on any combination of personal identifiers or any organizational codes. The Discrete Transaction File consists of an indexed file of one year's duration. At the end of the current year the data is stored on microfilm in a format basically the same as that of the original Discrete History File. While there was a provision in the original MIF for complete dates of birth and full given names, this information was not readily available for all individuals in computerized form when the file was first developed. As a result, the missing information was retrieved from approximately 15,000 paper files and transcribed into a machine-readable format; eventually ending up in the MIF.

The files which were initially combined to form the MIF were as follows:

1. The IDFILE composed of 205,325 records derived from paper files referring to registry participants from all years of the registry. In addition, the IDNEW file comprised of 33,368 records derived from paper files was added later.
2. The IDCOMB file composed of 35,729 records which was compiled during the period 1951 to 1963.
3. The INACLAM file dating from 1964 to 1977 comprised of 76,576 records.
4. The IDSIN file composed of 88,491 records having a S.I.N. number. These records were generated from 1977 onwards.

These files total 439,489 records. Duplicate records were created for individuals having a two-part surname. In these cases a record was added for each part of the surname so that the

individual could be searched under each name. The final number of records was 442,226. Figure 1 shows how these records were distributed in each phase of work on the NDR before linkage to the CMDB.

4. Major Stages of the Record Linkage Phase

I. Evaluating the Master Identification File
II. Correcting the Master Identification File
III. Internal linkage of the Master Identification File
IV. Merging of MIF data with dose history data.
V. MIF - Canadian Mortality Data Base linkage.
VI. Creation of the analysis file containing matched dose and death information.

Stages I-III are completed and are described in this report. Subsequent stages will be described in a future report. Figure 2 is a flowchart that delineates the necessary steps involved in completing the study.

B. METHODS

1. Stage I: Evaluating the Master Identification File

1.1 Contents (codes specific to NDR employment variables are indicated by an * and delineated in Appendix A)

(a) Surname

All records in the file have a surname. The length of the surname field is twelve characters. The following anomalies were found in the file:

(1) Many names had imbedded blanks. These could be due to two part names or to names having a prefix such as LA, DE, or VON. An especially common case of prefix was SR for nuns.

(2) Females often had one record for their maiden name and one record for their married name. Unless there was a common numeric identifier, these records remained separate.

(3) In some cases the original surname was longer than twelve characters which was the allocated field length. For these cases, BRMD had placed an asterisk at the end of the field. Duplicate records were included in the file for the part of the name that had been truncated.

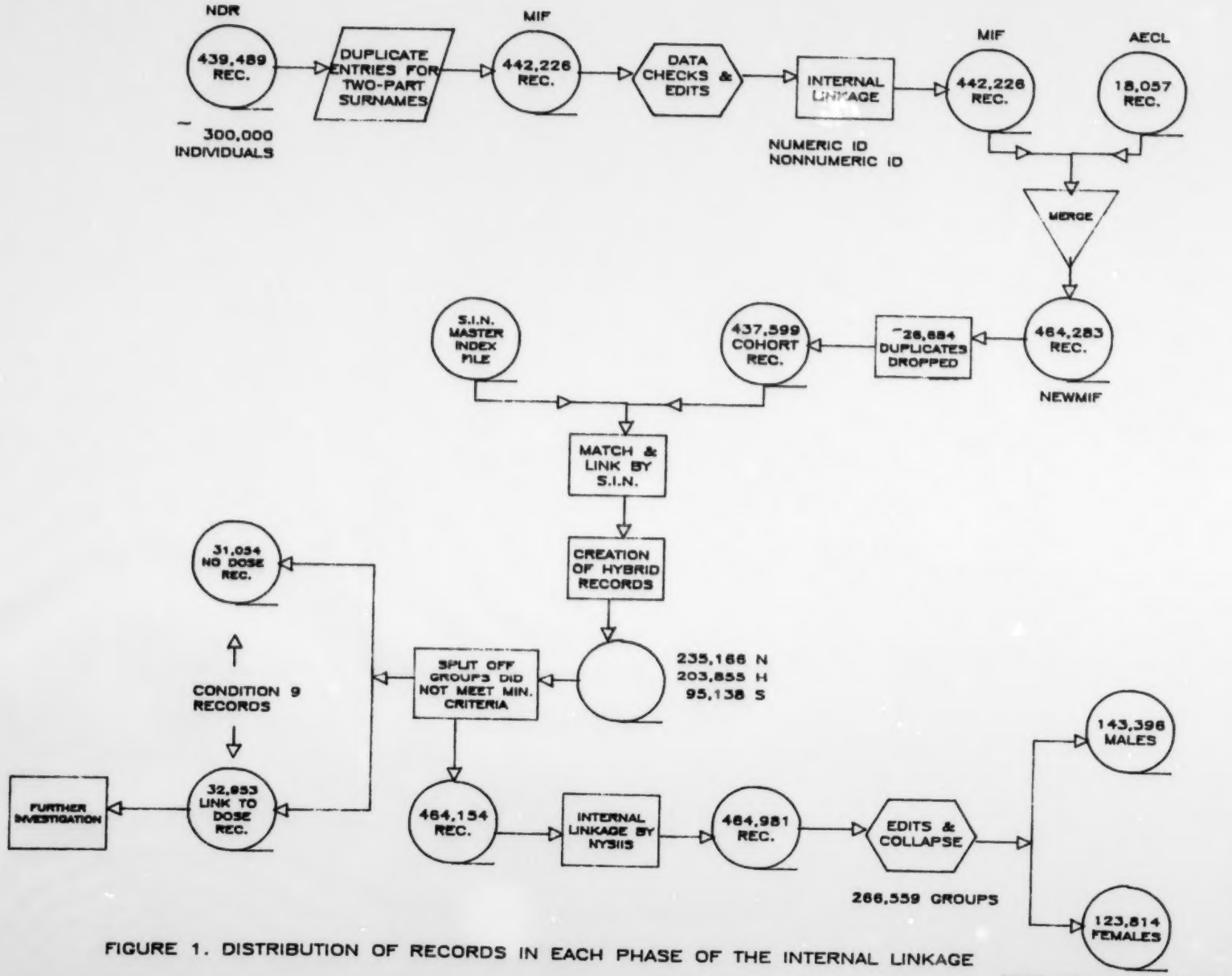

FIGURE 1. DISTRIBUTION OF RECORDS IN EACH PHASE OF THE INTERNAL LINKAGE

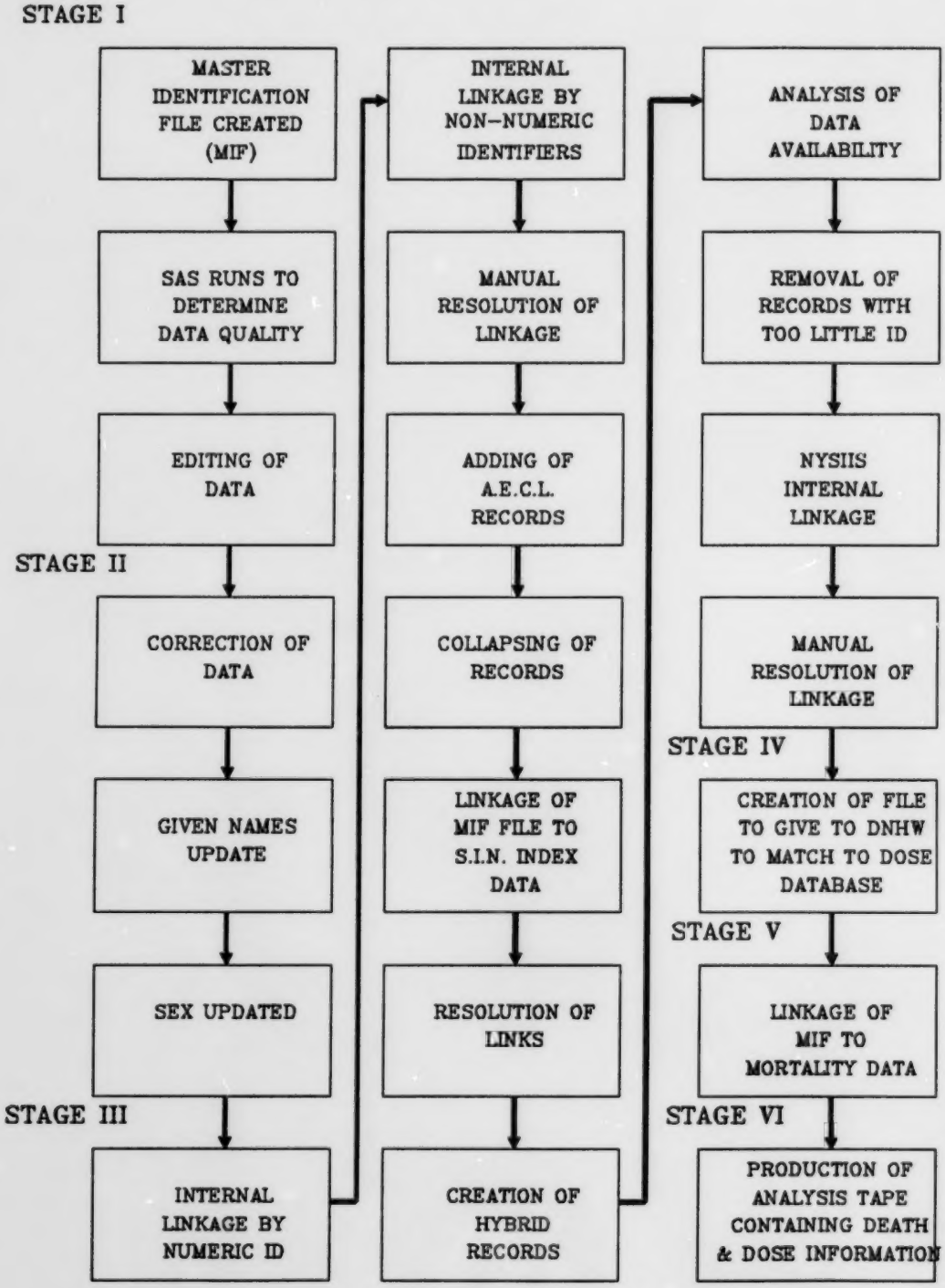
STAGE I
MASTER IDENTIFICATION FILE CREATED (MIF)
SAS RUNS TO DETERMINE DATA QUALITY
EDITING OF DATA
STAGE II
CORRECTION OF DATA
GIVEN NAMES UPDATE
SEX UPDATED
STAGE III
INTERNAL LINKAGE BY NUMERIC ID
INTERNAL LINKAGE BY NON-NUMERIC IDENTIFIERS
MANUAL RESOLUTION OF LINKAGE
ADDING OF A.E.C.L. RECORDS
COLLAPSING OF RECORDS
LINKAGE OF MIF FILE TO S.I.N. INDEX DATA
RESOLUTION OF LINKS
CREATION OF HYBRID RECORDS
ANALYSIS OF DATA AVAILABILITY
REMOVAL OF RECORDS WITH TOO LITTLE ID
NYSIIS INTERNAL LINKAGE
MANUAL RESOLUTION OF LINKAGE
STAGE IV
CREATION OF FILE TO GIVE TO DNHW TO MATCH TO DOSE DATABASE
STAGE V
LINKAGE OF MIF TO MORTALITY DATA
STAGE VI
PRODUCTION OF ANALYSIS TAPE CONTAINING DEATH & DOSE INFORMATION

FIGURE 2: OVERVIEW OF RECORD LINKAGE PHASE

(b) **Given names**
There are fields in the file for a first given name and a second given name. However many of the records only have initials, particularly those records from the IDSIN file. The other problem was that some given name fields contained two-part names.

(c) **Sex**
Most records had a valid sex code; however, some sex codes were assigned a code denoting unknown, and some of the sex codes were in error.

(d) **Date of Birth**
Fields for year, month and day of birth exist on the file, however a large proportion of records only had the year available (Table 1).

(e) **Social Insurance Number (S.I.N.)**
This number was present on about two thirds of the total file. Most of the numbers were valid. The S.I.N. was used as the numeric identifier from 1977 onwards.

(f) **Radiation Protection Bureau Number (RPB)**
This numeric identifier was assigned to individuals during the years 1973 to 1977.

(g) **Namecode**
This number was assigned to individuals during the years 1964 to 1972.

(h) **Combseq**
The Combseq number, for the purpose of this study, was a sequential number assigned to an alphabetically sorted list of individuals who had dose history records during the years 1951 to 1963, but had no other numerical identification.

(i) **Place of Birth**
The codes used to indicate the place of birth are consistent with those used at Statistics Canada. There is a code for every province and territory, and a code for each country.

(j) **Group Serial Number**
This code identifies the individual's employer.

(k) **Employer Classification**
This code is a broad classification of type of employer. There are 13 codes.

(l) **Job Class**
This code identifies the type of occupation. There are 80 different codes.

(m) Year of Entry
This indicates the year of enrolment (i.e. year first monitored).

(n) Year of Last Contact
This indicates the most recent year for which a dosimeter had been assigned and a dose reported for an individual.

(o) Province of Last Contact*
This indicates the province where the individual was last known to have been monitored.

1.2 Editing of Identifiers on the MIF

There were inconsistencies in coding and data entry in some of the fields in the MIF. For example, some names had characters other than letter A to Z in them, birth months greater than 12 etc. An edit program was written to determine the quality of the identifiers and to prepare a file of error records which could be corrected. All records were passed through the edit program. The error file contained approximately 12,000 records and was forwarded to the BRMD for correction. A tape of corrected records was subsequently returned to Statistics Canada, and the MIF was updated with the corrections. The details of the edit rules are presented in Appendix B.

1.3 Availability of Identifiers

After the edits and corrections, a number of computer runs using SAS were made to determine the availability of each identifiers in the MIF. The results in Table 1 provide the percentage of records which currently had the listed item. The identifiers marked with an asterisk are the minimum personal identifiers usually required for a CMDB linkage. It was anticipated that these percentages would increase when the file was collapsed from multiple records per individual to single records after the internal linkage.

2. Stage II: Correcting the Master Identification File (MIF)

Customized computer programs have been written to check discrepancies in sex coding and in given names. Where a discrepancy was identified it was reviewed by BRMD with updates being made by Statistics Canada. These programs are further described.

(a) Sex Code Updates
Not all records in the file had a sex code. A program was written which assigned sex codes by comparing the given name to a dictionary of names generated from the Mortality Data-

base. The dictionary gives the frequency of occurrence of given names by sex. If the frequency for one sex was at least twenty times greater than for the other then this sex code was assigned to the record. The number of new male codes assigned was 3,457 and the number of female codes was 4,344. In fact, all records were run through the sex coding program. Where sex was already assigned, but disagreed with the sex code derived from the program, the record was researched by BRMD staff and corrected if necessary.

Table 1: Percentage of Identifiers Available on the Master Identification File (after the edit program).

Variable	Percent availability
S.I.N.	70%
RPB Number	34%
Namecode	23%
*Surname	100%
*First Initial	98%
Second Initial	33%
*Rest of First Name	44%
Rest of Second Name	07%
*Sex	87%
*Birth Year	85%
*Birth Month	49%
*Birth Day	49%
Birth Place	12%
Year of Entry	93%
Year of Last Contact	93%
Job Classification	52%
Group Number	93%
Type of Employer	93%

* minimum identifiers usually expected for a CMDB linkage

(b) <u>Given Name Updates</u>

There were approximately 1,200 first given names on the MIF that were invalid. Some had invalid characters, others had embedded blanks, and others were two part names. A program was written which adjusted the way the first and second given names were coded according to the rules listed below.

(1) Where the first given name was in two parts separated by a blank, the second part of the name was moved to the second given name field providing it was blank. If the second given name field was not blank the first

given name was left as it was. Where the second given name field was the same as the second part of the first given name the latter was removed.

(2) Some names were in two parts; these were simply concatenated. (e.g. JO ANNE becomes JOANNE)

3. Stage III: Internal Linkage of the Master Identification File

One of the major tasks in this study is to create individual employment exposure histories. Records were fragmented for two main reasons: (1) there were multiple entries for persons who had left and rejoined the workforce at different times and, (2) name changes, especially in women.

To bring together fragmented records, it was necessary to compare the records within the file so that all those records belonging to the "same" individual could be assigned a common identifying number. This type of one-file comparison is normally referred to as an internal linkage.

Several internal linkages were done during the first three stages of the record linkage phase. The numeric linkage was done first as this type of linkage is cheaper and easier to perform. Details of each of these linkages follow.

3.1 Linkage By Numeric Identification Numbers

Several numeric systems have been used over the history of the NDR to identify individuals (see Section B, 1.1, (e) to (h)). These numeric identifiers are S.I.N., RPB, Namecode, and Combseq. Records could contain one or more of these numbers, or none of them depending on the source of the information and the year of registration. Table 2 gives the distribution of ID numbers among the original 439,489 records. The first internal linkage was performed to group together all records having like ID numbers.

(a) Preprocessing

For the internal numerical linkage, a special version of the original MIF was created by Statistics Canada which was composed of the original records, but with all identifying information removed except for ID numbers and a unique sequential record number. This file was expanded by creating a separate duplicate record for every different ID Number on the record (Table 3). Duplicate flags were assigned to all records generated from one original record with 0=one record, 1=two or more, etc..

Table 2: Distribution of Numeric Identification Numbers (0=absent, 1=present)

Numeric Identifiers				No. of records	Percent of total records
S.I.N.	RPB	Namecode	Combseq		
0	0	0	0	11,605	2.64
0	0	0	1	35,888	8.17
0	0	1	0	2,457	0.56
0	0	1	1	254	0.06
0	1	0	0	707	0.16
0	1	1	0	81,807	18.61
0	1	1	1	2,537	0.58
1	0	0	0	236,342	53.78
1	0	0	1	30	0.01
1	0	1	0	324	0.07
1	0	1	1	19	0.00
1	1	0	0	52,065	11.85
1	1	0	1	67	0.01
1	1	1	0	14,399	3.28
1	1	1	1	988	0.22
69.2%	34.7%	23.4%	9.1%	439,489	100.00

A pocket is a field(s) used to sort or subdivide records for comparisons where each record within a pocket is compared to each other. At this stage, to simplify the use of four ID numbers, a field called pocket-ID was created such that each ID number was prefixed with either a "1", "2", "3" or "4". These prefixes identified the ID number for which the particular record was created (1-S.I.N., 2-RPB, 3-Namecode, 4-Combseq). Use of the pocket-ID is shown in the expanded template of Table 3.

This special version of the original MIF was then sorted by the newly created pocket-ID to bring together, for comparison, all records having the same pocket-ID.

(b) <u>Compare Rules</u>

The compare rules for this phase of the project were very simple. If any of the ID numbers within a pocket agreed, a weight of 10 was assigned by GIRLS. A linkage weight of 10 or more was accepted as a good link.

Table 3: Record Layout for Expanded File Used During Preprocessing

Original Template (example)

Recno.	Names & Other Fields	S.I.N.	RPB	NC	Combseq
246	----------	X			
247	"		Y	Z	
248	"	X	W	V	
249	"				U

Expanded Template (example, see text for definition of pocket-ID and use of prefixes)

Recno.	Dupflag	1 S.I.N.	2 RPB	3 NC	4 Combseq	Pocket-ID
246	0	X				1-X
247 247	1 2		Y Y	Z Z		2-Y 3-Z
248 248 248	1 2 3	X X X	W W W	V V V		1-X 2-W 3-V
249	0				U	4-U

During the GIRLS grouping operation a common group number was assigned to all records which linked together on any of their ID numbers. After the linkage a SAS program was written which took the group number generated by the linkage and added it to the appropriate record in the original MIF. Records which had no ID numbers, and records which had unique ID numbers were assigned a group number of 0.

3.2 Linkage By Numeric Identification Using Non-Numerical Identifiers

Each record with the same group number was then internally "linked" within the group using the non-numerical personal identifiers (ie. surname, given name, birth date), to check that different persons had not been grouped together because of an error in one or more of the MIF ID numbers. Only records from the MIF having a non-zero group number were included in this linkage. There were 142,778 of these records.

(a) Preprocessing the File

The files containing the records used in the linkage needed to be prepared for input into the GIRLS system. Preprocessing involved the tasks outlined below.

(1) Standardization of Surnames

The surname fields were corrected. All non-alphabetic characters were removed from names. Where a blank character divided the surname, alternate records were generated for each part. If the second part of the name only had one character it was removed. If the first part of the name was a standard prefix then no alternate record was created and the two parts of the name were concatenated. A new field was created for the modified surname and the old version was retained.

(2) Encoding of Fields

To reduce the problems arising out of variant spellings, a phonetically coded surname sequence (NYSIIS code) was added to each record on the MIF. Weight-class codes were generated. These codes are determined by the commonality of the surname and they indicate the linkage weight which will be assigned during the compare of surnames.

(3) Creation of Duplicate Records

Duplicate flags were assigned to all records generated from the same original record. (i.e. where two records were created for a two-part surname). Duplicate records in the MIF have the same sequence number.

(b) Pocket Identifier

For this linkage, the pocket identifier was the assigned group number from numerical linkage described in 3.1 (i.e. only records which had some form of common ID number were compared together).

(c) Compare Rules and Weight Sets

The actual rules and weights used during this linkage are listed in Appendix C. Weights for given names were derived from the MIF. They were created for males and females separately for agreement on 1,2,3,4 and 7 characters, and cross agreement.

The following data items were compared during this linkage:

a. Birth year
b. Birth month
c. Birth day
d. Initials and cross initials
e. First given name
f. Second given name
g. Cross agreement of given names
h. NYSIIS and surname
i. Birth place
j. Year last known alive
k. Job class
l. Year of entry
m. Last known province of monitoring
o. Employer class

(d) Production Runs

The threshold value, the value above which all links are taken as correct, was set extremely low, at -999, so that all possible links within a group were accepted for further resolution and analysis. The production runs were submitted in three parts because of the limited amount of space available and the size of this file (same rules and thresholds, different file size).

(e) Manual Resolution

Listings of the results of the linkage were given to the BRMD to determine which links to reject and which to accept. Links above a specified weight were accepted. A printout of all links below the specified weight was sent to the BRMD to be verified by their staff by referring to their LDHS database. Any required changes were indicated on the listings and returned to Statistics Canada.

The main difficulty encountered during this process was with ID numbers. Sometimes two different values of the same ID number were assigned to the same individual, or the same number was assigned to more than one person. Namecode was the worst offender, followed by S.I.N., then RPB. Many of these errors were corrected by reference to old files at the BRMD.

A further problem was conflicting sex codes between records in the group. This was resolved by accepting the code occurring most often. There were 19,376 groups resolved in this manner. The computer updates to modify the linkages were made at Statistics Canada.

(f) Determining Distribution of Group Size

Table 4 shows the results of a SAS run written to determine the distribution of group size, at the completion of this linkage after the above adjustments were made to the file.

Table 4: Distribution of Records per Group

Size of group (No. of records)	Number of groups
2	39,539
3	11,140
4	4,332
5	1,299
6	488
7	178
8	67
9	30
10	14
11	8
12	5
13	0
14	2
	57,105

3.3 Adding Atomic Energy of Canada Limited Records to the MIF

Atomic Energy of Canada Limited (AECL) did not send its dose records to the NDR on a yearly basis until 1981. A file containing 18,057 pre-1981 records from AECL was given to NDR to link to the already existing MIF. This file contained the historic AECL records to 1972 which were now in machine readable form, and the dose tapes from 1973 to 1980.

Figure 3 illustrates the steps taken to add these records. First of all, the records in the AECL file were assigned a record identification number. The file was then split in two; one part containing records with a S.I.N. and the other part having records without a S.I.N. The AECL records with a S.I.N. were matched to the MIF records by S.I.N. Matching

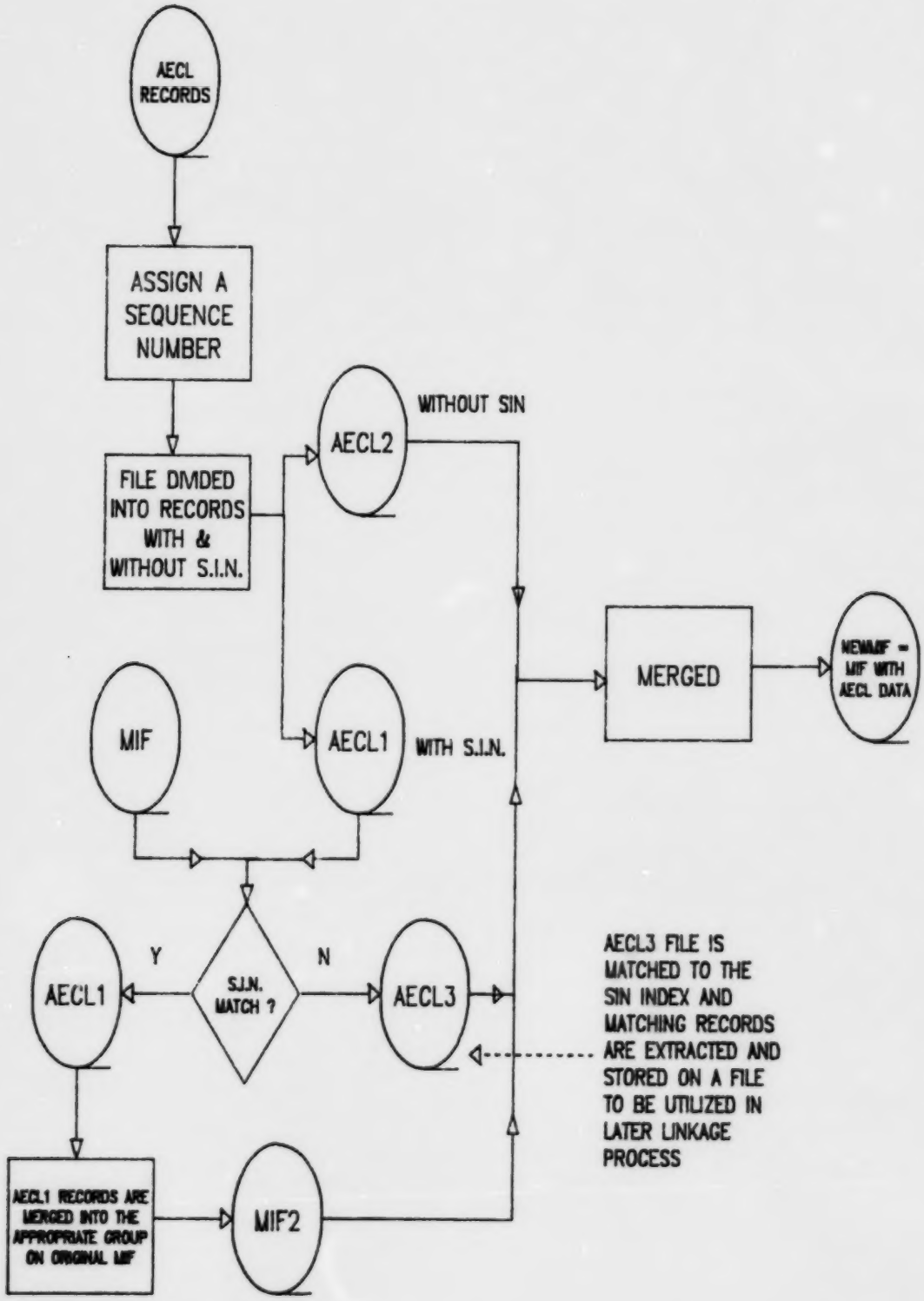
AECL RECORDS
ASSIGN A SEQUENCE NUMBER
FILE DIVIDED INTO RECORDS WITH & WITHOUT S.I.N.
AECL2
WITHOUT SIN
MERGED
NEW MIF - MIF WITH AECL DATA
MIF
AECL1
WITH S.I.N.
AECL1
Y
S.I.N. MATCH ?
N
AECL3
AECL3 FILE IS MATCHED TO THE SIN INDEX AND MATCHING RECORDS ARE EXTRACTED AND STORED ON A FILE TO BE UTILIZED IN LATER LINKAGE PROCESS
AECL1 RECORDS ARE MERGED INTO THE APPROPRIATE GROUP ON ORIGINAL MIF
MIF2

FIGURE 3: ADDING AECL RECORDS

records were merged into the MIF and assigned the appropriate group number. The unmatched AECL records (including those with no S.I.N.) and the updated MIF were combined to create a new identification file (NEWMIF). There were 464,283 records on this new file.

3.4 Collapsing the NEWMIF Records

Many of the groups in the NEWMIF contained records that were virtually the same. It was decided to combine records having no conflicting identification into a single record so that superfluous records could be deleted. A program was written to compare records within one group, to move personal identifiers from one record to another where appropriate, and to remove duplicate records.

The rules for merging the identifiers were as follows:

(1) The given name was moved from non-blank to blank fields or extended if a longer compatible name was present. i.e. R, Ron, and Ronald all become Ronald if Ronald appeared on one of the records in the group.

(2) Birth day, birth month, birth year were moved into blank fields providing there were no conflicts in other records in the same group.

(3) The identity numbers - S.I.N., RPB, Namecode, and Combseq - were moved into blank fields if there were no conflicts.

(4) Place of birth was moved into a blank field if there was no conflict.

(5) Surname, year of entry, year of last contact, province of last contact, job classification, and employer classification were moved into blank fields if there were no conflicts.

After merging the data as described above, if a pair of records were identical (i.e. agreed on surname, given names, birth place and ID numbers), one record was dropped from the file and one was retained. The number of records dropped was approximately 26,684, leaving 437,599 records.

3.5 Master Identification File - Social Insurance Number Index File Linkage

After the reduction of the NEWMIF by amalgamating data from matching records and eliminating duplicates, programs were run to reassess the availability of data for the CMDB linkage. At this stage, the NEWMIF consisted of 437,599 records. These records correspond to a total of 371,849 possible individuals. The distribution of the number of linkable fields missing per individual is presented in Table 5.

Table 5: Missing Fields After Collapsing NEWMIF Records

Number of fields missing per group	Frequency of occurrence	Percent occurrence
0	167,066	44.9
1	12,522	3.4
2	46,274	12.4
3	9,549	2.6
4	3,686	1.0
5	31,952	8.6
6	60,157	16.2
7	30,754	8.3
8	9,889	2.7
	371,849	100.0

These results indicated that over 50% of the NEWMIF file had missing information thus making the NYSIIS internal linkage quite difficult. It was decided to use the S.I.N. Index so that individuals could be identified unambiguously. Where a NDR record with missing information linked to a S.I.N. Index record, a "hybrid" record could be formed containing information from both files. A better linkage to the CMDB could be achieved using this "hybrid" record.

Of the 437,599 records on the NEWMIF at this time, 294,895 records had a non-blank S.I.N. field. Of these 230,020 were unique, 13 individuals had two different S.I.N. values, 821 S.I.N. values were shared by two different individuals, and the remainder were duplicate entries.

(a) Preprocessing

In preparation for the S.I.N.- NEWMIF linkage, the relevant records on the S.I.N. Index file were extracted by matching

the NEWMIF to the S.I.N. Index, by S.I.N. Of these S.I.N. records extracted, 229,906 had unique S.I.N.s. The file of S.I.N. records was then linked to the NEWMIF using the GIRLS record linkage system with personal identifiers since a match on S.I.N. alone does not guarantee a true link.

The data items used in the S.I.N.- NEWMIF linkage included: birth year, month and day, sex, given names, and surname. The threshold value for the acceptance of a link was set at -100. The detailed weights are provided in Appendix D.

(b) Production Run

The production run was done in five parts to cope with the large file sizes. The results of each part were combined and all links having a weight less than 60 (2,334) were printed out for visual inspection.

(c) Manual Resolution

The decision as to whether the links were true or false was left to the judgement of the person doing the manual resolution at Statistics Canada. However, the following rules were generally observed.

1. Different surnames in women were allowed.
2. Blank sex and birth date were allowed.
3. If initials agreed, and sex agreed,then birth date was allowed to disagree.
4. If initials, month and day of birth agreed, then birth year could disagree.
5. Cross agreement of given name or initial was allowed if the year of birth agreed.

Links were categorized into "true", "possible" or "false". If the agreement between a pair of records was not sufficient for a "true" link, then the status of "possible" link was allowed if the agreements warranted this. The number of links updated after visual inspection was 465. Codes were assigned to indicate the link category; "true"=1, "possible"=2. After the manual resolution and computer updates were completed, the results of the linkage revealed that of the original records extracted, 260,504 found a "true" link to the NEWMIF and 989 were "possible".

(d) Creation of Hybrid Records

In an attempt to maximize the identifying information and decrease the number of records in the NEWMIF file, information from the NDR record was combined with its matching S.I.N. Index record to create "hybrid" records. This was

done only if the linkage was true and according to the following rules.

1. Given names were moved to blank fields or to fields where there was only an initial which agreed with the S.I.N. name.
2. Birth day, birth month, and birth year were moved into blank fields.
3. Sex was moved into a blank field.

At the end of this process, the S.I.N. record was dropped if there was either a hybrid record or an NEWMIF record with the same identifying data. The S.I.N. records involved in possible matches were kept on the merged file with a status flag. In addition, a "record type" flag was added to each record indicating whether it was a hybrid record(H), an unaltered NDR record(N), or a S.I.N. record(S). After the merging of the S.I.N. file and the NEWMIF, there were 203,-855 hybrids, 95,138 S.I.N. records that had found matches but could not be used to create hybrid records, and 235,166 unchanged NDR records.

(e) **Availability of Identifiers**

The new file was evaluated for availability of various identifiers. Records were classified according to four conditions as follows:

1= Records containing surname, birth year and full first given name.
2= Records containing surname, first initial, and full birth date.
3= Records containing surname, first initial, and birth year.
9= None of the above (i.e. less than the above).

Table 6 displays the breakdown of the 534,159 records that resulted from merging the S.I.N. Index data with the NEWMIF distributed across the four conditions described above.

From the NEWMIF, 64,007 groups (individuals) were split off because they did not have the minimum items required to participate in the mortality linkage (surname, first initial) and in the analysis (birth year,sex). These records are referred to as "condition 9" records. From Table 6, it can be seen that some of these records were linked with the S.I.N. Index because "hybrid" records had been created. If, for example, the NEWMIF record had a S.I.N., surname, and first initial but did not have birth year and/or sex code, it would have been

Table 6: Distribution of NEWMIF by Record Type (H=hybrid, N=NDR, S=S.I.N.)

Condition	Record type	Number of records
1	H	186,371
1	N	99,251
1	S	94,983
2	H	6,187
2	N	4,623
2	S	155
3	H	1
3	N	72,583
9	H	11,296
9	N	58,709

deemed a condition 9 record even if it found a true match with the S.I.N. Index because identifier information from the S.I.N. Index cannot be released outside Statistics Canada. The records currently held in the condition 9 file require further checking before carrying out any mortality search.

A tape file of all condition 9 records was sent to BRMD for further examination. A total of 31,054 groups were dysfunctional in that they had no numerical keys or no dose records. Preliminary examination of the remaining 32,953 groups, that apparently link to dose records, shows that some are already in NEWMIF. This has been determined by an examination of rare surnames and of yearly dose records. For those not already in NEWMIF, more identifying information may be available later so that they can eventually be included in the study. At the present, they will not affect the number of individuals in the current cohort for the mortality linkage. At most, these groups account for only 3% of the collective dose in the NDR.

(f) Adding of Pseudo-SINs

The Combseq number was not carried over to the new LDHS when it was set up in 1986. Therefore, it was decided at this stage to remove the Combseq number from NEWMIF and replace it with a number called a PSEUDO-SIN located in the S.I.N. field. This would make the NEWMIF compatible with the LDHS database. To achieve this, a tape was created at Statistics Canada containing all records, mostly AECL records, from the

NEWMIF which contained a Combseq number. The tape was sent to the BRMD, who subsequently returned a tape with the same records and a PSEUDO-SIN in the S.I.N. field. These PSEUDO-SINs were then merged into the NEWMIF.

3.6 Linkage of the NEWMIF By NYSIIS Code

The last stage of the internal linkage of the NEWMIF was a linkage involving all the records on the current file which now contained original NDR records, hybrid records and S.I.N. records. The pocket identifier for the linkage was the NYSIIS code of the surname. The file now contained 464,154 records (condition 9 groups were removed). This final internal linkage will result in bringing together all fragmented records relating to the same individual. The final results of this will serve as input to the mortality linkage.

The details involved in this NYSIIS internal linkage were as follows:

(a) Preprocessing the NEWMIF Records

(1) The 764 records which referred to nuns were duplicated and put in a special pocket which was identified by creating an artificial NYSIIS code called SOEUR. A special pocket was created because the records for nuns have special linkage problems in that their religious names are very similar.

(2) Any nonconforming surnames were corrected by removing blanks and illegal characters.

(3) The weight class codes were generated for the surnames. Three different codes were assigned to each record. One code was for agreement on NYSIIS and seven letters of the name, another was for agreement on NYSIIS and four letters of the name and one for agreement on NYSIIS only. Each code is associated with a weight set which is generated from the CMDB. The magnitude of the weights are in inverse proportion to the frequency of occurrence of the name.

(4) The birth years were converted to binary format.

(5) Invalid values of birth month and day were removed.

(b) Compare Rules

The identifiers used in the compare phase were the birth year, the birth month, the birth day, the first and second

given names, the surname, the birth place, the year last known alive, the job class, the year of entry to the registry, the last known province of employment, and the identification numbers. The details of the comparison rules are given in Appendix E.

(c) Production Runs

The NEWMIF was sorted by the alphabetical sequence of the NYSIIS code. It was then divided into six separate versions (in the context of the GIRLS system, a version is a subset of the original file) since all the records could not be accommodated in one production run. A seventh version contained the duplicate records which pertained to nuns. The distribution of the records in each version are listed in Table 7.

Table 7: NEWMIF Record Distribution for NYSIIS Production Runs

Version	NYSIIS code range	Number of records compared
1	A-CAP	84,009
2	CAQ-F	76,692
3	G-LAS	91,963
4	LAT-RAK	105,889
5	RAL-SOET	40,755
6	SOEU-ZZZ	64,902
7	SOEUR	764

When choosing threshold values a balance has to be made between the number of poor links accepted by the system and the risk of losing good links. It is important to note that although different thresholds were used, it does not mean that records would be dropped from the linkage with the CMDB. At most, it would mean that more single records would advance to the mortality linkage.

In order to choose appropriate lower threshold values, samples of linked records at different weights were visually inspected prior to the final production runs. Following this inspection, the first version was run at a lower threshold of 51. This value worked quite well, but it was decided to try a lower value of 41 for the second version to see if there was any significant difference in the number of true links.

In retrospect, the value of 41 was too low. Although few, if any, true links were not grouped into the appropriate

individual history far too many bad links were included which lead to excessive manual resolution. For versions 3,4,5 and 6, the lower threshold was raised to 63. After examination of the results of versions 3 and 4, it was felt that 63 was the most appropriate threshold to use through to completion of the production runs to eliminate most spurious links. The special version of nun records was run using 51 as the lower threshold.

(d) Manual Resolution

Manual resolution was carried out using a printout of all groups containing a link in the "doubtful" range i.e. between a weight of 119 to the lower threshold value. The lower threshold values used and the distribution of groups and records output for each version prior to manual resolution are shown in Table 8.

Table 8: Distribution of NYSIIS Groups Prior to Manual Resolution

Version	Lower threshold values	Number of groups in output	Number of records in output
1	51	5,578	14,511
2	41	5,472	13,265
3	63	6,434	15,190
4	63	7,541	18,143
5	63	2,916	6,943
6	63	4,814	11,407
7	51	134	457

Manual resolution was done in two stages. The first stage was a visual inspection. Each of the groups was carefully examined. If no changes had to be made to a group then it was left unmarked. If there were links that were unacceptable, the group was broken up into subgroups or single records. This was done by marking the computer output showing how a group was to be divided. When a decision on a group could not be made simply by visual inspection, it was marked for resolution to be done using the information stored on LDHS.

To expedite the manual resolution process, a computer terminal was installed at Statistics Canada which allowed BRMD staff immediate access to the LDHS at DNHW. The dose histories of records involved in doubtful links were accessed from the LDHS and inspected to determine if one or more

individuals were involved in the assigned group. This information was useful in deciding whether or not to reconfigure the groups in question.

Computer updates were made as required to effect the necessary changes to the doubtful groups. The distribution of the links after manual resolution and the computer updates are displayed in Table 9. It should be noted that links in the context of Table 9 are not equivalent to records. For example, a group containing 4 records has the possibility of generating 6 links when internally linked. Also note that resolved links are included in true or rejected links.

Table 9: Distribution of Links After Manual Resolution

Version	Number of true links	Number of rejected links	Number of links resolved
1	10,849	44,743	5,060
2	10,000	27,312	4,943
3	11,062	42,682	2,510
4	13,402	38,898	2,963
5	5,067	27,088	2,916
6	8,245	26,183	2,135
7	324	3,121	401

(e) Merging of the Versions

After the updates to each version had been made, all the versions were merged together on a single file. One individual could have one or several records, but all records which refer to the same individual now have the same sequence number. There were 464,974 records divided among 266,465 groups (or individual sequence numbers). Subsequently, there were a few more edits performed, records were collapsed, and 94 new groups added. There were 651 records with no sex code available. The final NEWMIF cohort file contained 392,454 records relating to 266,559 individuals. There were 143,396 males and 123,814 females available for the mortality linkage. The details of the mortality linkage are addressed in a separate second report.

4. Stage IV: Merging of MIF Data with Dose History Data

A file was created for the BRMD which would allow for linking their original MIF records to the LDHS database. The records on

the file contained the Statistics Canada sequence number which now identifies the individual, and all ID numbers which refer to the individual and the surname. The data given to the BRMD only consisted of the data originally provided by them. S.I.N. Index file data were not released. In order to create the file with NDR data only, the records from the file output from the NYSIIS linkage were matched to the MIF prior to merging with S.I.N. data. This file contained the condition 9 records. There were 439,021 records on this file. The BRMD will use this file to access all the relevant annual dose records for an individual and to create a composite dose history for each individual. After these reviews, a computer tape of individual dose histories will be created.

5. Future Work - Stages V and VI

Individual dose histories will be merged with the results of the mortality linkage to form an analysis file. Details regarding Stages V and VI are described in a separate second report.

REFERENCES

1. National Research Council, Committee on the Biological Effects of Ionizing Radiation (BEIR V). Health Effects of Exposure to Low Levels of Ionizing Radiation. Washington, D.C.: National Academy Press, 1990 (and other BEIR Reports).

2. United Nations Scientific Committee on the Effects of Ionizing Radiation (UNSCEAR). Sources, Effects and Risk of Ionizing Radiations. UN Publication, N.Y., 1988 (and other UNSCEAR Reports).

3. ICRP Publication 26. Recommendations of the International Commission on Radiological Protection. Annals of the ICRP, 1, No. 3, 1977 (and other ICRP Publications).

4. NCRP Report, No. 64. Influence of Dose and Its Distribution in Time on Dose-response Relationships for Low-LET Radiations. NCRP, Washington, D.C., 1980 (and other NCRP Reports).

5. Protocol for the National Registry for Radiation Workers. (SC Darby, ed.). National Radiological Protection Board, Harwell, U.K.,1981, pp.1-24.

6. Cardis E, Kaldor J. Protocol of Combined Analysis of Cancer Mortality Among Nuclear Industry Workers. Internal Report 89-005, IARC, Lyon, 1989.

7. Fair M.E. Radiation Studies in Canada: The National Files and Facilities Needed. In: Radiation Protection: Past and Future. Atomic Energy of Canada Limited, Report AECL - 9959 (1989).

8. Land CE. Estimating Cancer Risk from Low Dose Ionizing Radiation. Science 1980;209:1197-1203.

9. Ashmore JP, Grogan D. The National Dose Registry for Radiation Workers in Canada. Radiation Protection Dosimetry 1985; 11(2): 95-100.

10. Ashmore JP, Davies BF. The National Dose Registry: A Centralized Record Keeping System for Radiation Workers in Canada. Seminar on the Application of Computer Technology to Radiation Protection. International Atomic Energy Agency, June 1987.

11. Smith ME, Newcombe HB, Use of Canadian Mortality Data Base for Epidemiologic Follow-up. Canadian Journal of Public Health 1982; 73: 39-46.

12. Smith ME, Silins J, Generalized Iterative Record Linkage System. In: Social Statistics Section, Proceedings of the American Statistical Association, 1981: 128-137.

13. Occupational Radiation Exposure and Mortality: A Protocol for an Analysis Based on the National Dose Registry of Health & Welfare Canada; Health Protection Branch, Health & Welfare Canada, Ottawa. 1991. (In preparation).

14. Hill T. Generalized Iterative Record Linkage System GIRLS STRATEGY Release 2.7. Report from Research and General Systems, Informatics Services and Development Division, Statistics Canada. Tunney's Pasture, Ottawa, 1988.

15. Newcombe HB, Handbook of Record Linkage. Oxford University Press, 1988.

16. Fair ME, Newcombe HB, Lalonde P. Improved Mortality Searches for Ontario Miners using Social Insurance Index Identifiers. Report for Atomic Energy Control Board, Ottawa, 1988.

ACKNOWLEDGEMENTS

The extensive contributions of time, expertise and resources made towards the design of the study (in particular, the design and implementation of the record linkage phase) and the review of this report by the following organizations and individuals are gratefully acknowledged.

Provincial Vital Statistics Departments
Department of Employment and Immigration
The Laboratory Centre for Disease Control

Dr. Howard Newcombe - Consultant, and AECL (retired)

Dr.Pat Ashmore - Bureau of Radiation & Medical Devices (DNHW)
Brian Davies - Bureau of Radiation & Medical Devices (DNHW)

Guy Quesnel - Statistics Canada
Pierre Lalonde - Statistics Canada
Mary Werner - Statistics Canada

APPENDIX A

CODES FOR EMPLOYMENT-RELATED VARIABLES SPECIFIC TO NDR

PROVINCE OF LAST CONTACT : This is the province where the individual was last monitored.

CODE	PROVINCE
0	Newfoundland
1	Prince Edward Island
2	Nova Scotia
3	New Brunswick
4	Quebec
5	Ontario
6	Manitoba
7	Saskatchewan
8	Alberta
9	British Columbia
A	Northwest Territories
B	Yukon
C	Labrador
D	Other Countries

EMPLOYMENT CLASS : This variable is available from 1964 to the present.

CODE	CLASS
1	Dentist
2	Hospital
3	Industry
4	National Defence
5	Mobile X-Ray Units
6	Private Practitioners, Government, Vets, Clinics
7	Other Agencies, Research
8	Special Projects
9	Civil Defence
A	Uranium Mine
B	Non-Uranium Mine
C	Nuclear Power Station
D	Nuclear Fuel Refining Facility

JOB CLASSIFICATION

ADMINISTRATIVE:

CODE	CLASS
0001	Administrator
0002	Office Staff
0003	Safety Officer
0004	Other

MEDICAL:

CODE	CLASS
0010	Chiropractor
0011	Dentist
0025	Dental Hygienist
0012	Gynaecologist
0013	Laboratory Technician
0014	Medical Physicist
0015	Nurse
0016	Physician
0017	Radiological Technician (Diagnostic)
0018	Radiological Technician (Therapeutic)
0019	Radiologist (Diagnostic)
0020	Radiologist (Therapeutic)
0021	Veterinarian
0022	Ward Aide or Orderly
0024	Isotope Technician
0023	Other

INDUSTRIAL:

CODE	CLASS
0040	Dial Painter
0041	Instructor
0042	Instrument Technician
0043	Laboratory Technician
0044	Oil Logger
0045	Radiography
0046	Scientist and Engineer (Field)
0047	Scientist and Engineer (Laboratory)
0048	Other
0071	Salesman
0072	Student
0073	Visitor
0074	Other (Misc)

NUCLEAR REACTOR PERSONNEL: This breakdown was implemented in 1977. Prior to that time there was a single code "reactor worker.

CODE	CLASS
0510	Administration/Security/Janitorial
0520	Health Physics
0530	Chemical and Radiation Control
0540	Electrical Maintenance
0550	Mechanical Maintenance
0560	General Maintenance
0570	Fuel Handling
0580	Reactor Operations
0590	Control Technicians
0500	General- Scientific/Professional Staff
0598	Construction
0599	Visitors
0597	Training Staff
0545	Industrial Radiographers
0596	Summer Student

URANIUM MINES: This section is not fully implemented as not all mines provide the breakdown; however, all are divided by as uranium mine or non-uranium mine.

CODE	CLASS
0601	Underground Personnel
0602	Surface Personnel
0603	Visitor
0604	Office Staff
0610	Underground Miner
0615	Nurses
0620	Support Workers
0630	Underground Maintenance
0640	Electrician
0650	Surface Miner
0660	Surface Support Workers
0670	Surface Maintenance
0680	Mill Maintenance
0690	Mill Workers

NON-URANIUM MINES:

CODE	CLASS
0800/01	Underground Personnel
0802	Surface Personnel
0803	Visitor
0804	Office Staff
0810	Underground Miner
0815	Nurses
0820	Support Workers
0830	Underground Maintenance
0840	Electrician
0850	Surface Miner
0860	Surface Support Workers
0870	Surface Maintenance
0880	Mill Maintenance
0890	Mill Workers

GROUP SERIAL NUMBER

This number is unique to the employer; therefore the options are too numerous to list.

APPENDIX B

QUALITY OF DATA (1.2) EDIT RULES

1. The surname must have letters A-Z and no embedded blanks.
2. Given names must have letters A-Z and no blanks.
3. Sex must be 'M' or 'F' or blank.
4. Birth Day must be equal to '01-31' or blank.
5. Birth Month must be between '01 or 12' or blank. Sometimes the birth month and day will be interchanged.
6. Birth place should be blank or one of the codes used by Statistics Canada to denote province or country.
7. S.I.N. if present has to conform to the validity check.
8. RPB number, if present, has to be within the valid range of 1000198 to 1184831.
9. Namecode number has to be within the valid ranges. There are two types of Namecode, an old one and new one. The old one has values between *020000-*805000, the new code has values between 100010-985405 or 000000-032223.
10. Employer group number has to be in the valid range, 0001-20,000.

APPENDIX C

COMPARE RULES: INTERNAL LINKAGE BY NUMERICAL IDENTIFICATION USING NON-NUMERICAL IDENTIFIERS

Birth Year The year of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	55	- 2	53
Partial 2	differ by 1 yr	42	- 35	7
Partial 3	differ by 2-3 yrs	32	- 45	- 13
Partial 4	differ by 4-6 yrs	27	- 50	- 23
Partial 5	differ by 7-9 yrs	30	- 70	- 40
Partial 6	differ by 10 yrs	45	- 70	- 25
Missing	blank, either record		- 10	- 10

Birth Month The month of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	36	- 3	33
Partial 2	differ by 1 month	27	- 55	- 28
Partial 3	differ by 2-3 months	20	- 55	- 35
Partial 4	differ by 4-6 months	20	- 40	- 20
Partial 5	differ by 11 months	22	- 45	- 23
Disagree	all other cases	0	- 22	- 22
Missing	blank, either record		- 5	- 5

Birth Day The day of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	50	- 3	47
Partial 2	differ by 1 day	40	- 50	10
Partial 3	differ by 3 days	30	- 45	- 15
Partial 4	differ by 6 days	25	- 60	- 35
Partial 5	differ by 30 days	10	- 60	- 50
Disagree	all other cases		- 40	- 40
Missing	blank, either record		- 5	- 5

Birth Place

The birthplace codes are compared. The codes can be of a province or a country. Some records just have a code denoting Canada.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	Prov & Prov	30	- 2	28
Partial	Province & Canada	30	- 20	10
Disagree	all other cases		- 30	- 30

Initials of the First and Second Given Names

All combinations of the initials are compared e.g. first initials together, second initials together and first with second initials.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
	4 initials			
Full	2 agree G1=G1, G2=G2	85	- 15	70
Partial 2	2 agree G1=G2, G2=G1	90	- 60	30
Partial 3	G1=G1 or G2=G2	40	- 55	- 15
Partial 4	G1=G2 or G2=G1	45	- 80	- 35
Partial 5	all disagree	15	- 80	- 65
	3 Initials			
Partial 6	G1=G1	50	- 25	25
Partial 7	G1=G2 or G2=G1	50	- 55	- 5
Partial 8	all disagree	15	- 75	- 60
	2 Initials			
Partial 9	agree	60	- 35	25
Partial 10	disagree	35	- 80	- 45
Missing	blanks either		- 15	- 15

First Given Name

Assign level of agreement for the remaining characters of the first given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Second Given Name

Assign level of agreement for the remaining characters of the second given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

First Given Name to Second Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the first given name from File A to the second given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 17
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Second Given Name to First Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the second given name from File A to the first given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	29	- 2	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 54	5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Surname
Assign level of agreement for the NYSIIS code and surname. Thirteen characters are compared, six for NYSIIS, seven for the surname. Partial agreements occur when not all thirteen characters agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	13, NYSIIS & Surname	0	- 1	- 1
Partial 2	10, NYSIIS & 4ch. Surname	0	- 40	- 40
Partial 3	6, NYSIIS agree	0	- 60	- 60
Disagree	all other cases		- 75	- 75

Weight Class
Specifies a range of weights which depend on how common or rare the surname is. Higher values are given to rare names. The rules for surname and weight class are used together with the surname providing the negative component and the weight class the positive component.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	Very Common	70	-10	60
Partial 2	Common	70	-18	52
Partial 3	Relatively rare	90	-19	71
Partial 4	Rare	105	-30	75
Partial 5	Very rare	130	-30	100

Year Last Known Alive
The year last known alive on each record was compared.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	27	- 37	10

Job Class The job class on each record is compared.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	35	- 5	30
Disagree	all other cases		- 10	- 10

Year of Entry to Registry
The year of entry to the registry is compared and given a low weight for agreement.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	40	- 11	29
Disagree	all other cases		- 10	- 10

Last Known Province
The province where the person was last known to work was used for comparison.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	20	- 5	15
Disagree	all other cases		- 20	- 20

Employer Class
The type of employer of the person was used for comparison.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	25	- 5	20
Disagree	all other cases		- 15	- 15

APPENDIX D

COMPARE RULES FOR MIF - S.I.N. INDEX LINKAGE

File A - NDR MIF **File B** - S.I.N. Index Master File

Birth Year
The year of birth was compared between File A and File B

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	75	- 2	73
Partial 2	differ by 1 yr	63	- 30	33
Partial 3	differ by 2-3 yrs	55	- 49	6
Partial 4	differ by 4-6 yrs	48	- 53	- 5
Partial 5	differ by 7-9 yrs	40	- 70	- 30
Partial 6	differ by 10 yrs	30	- 50	- 20
Disagree	all other cases		- 70	- 70
Missing	blank, either record		10	10

Birth Month
The month of birth was compared between File A and File B.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	36	- 1	35
Partial 2	differ by 1 month	27	- 55	- 28
Partial 3	differ by 2-3 months	19	- 60	- 41
Partial 4	differ by 4-6 months	18	- 69	- 51
Partial 5	differ by 11 months	23	- 65	- 42
Disagree	all other cases		- 42	- 42
Missing	blank, either record		- 1	- 1

Birth Day
The day of birth was compared between File A and File B.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	51	- 3	48
Partial 2	differ by 1 day	42	- 62	- 20
Partial 3	differ by 3 days	30	- 60	- 30
Partial 4	differ by 9 days	26	- 60	- 34
Partial 5	differ by 30 days	7	- 51	- 44
Disagree	all other cases		- 40	- 40
Missing	blank, either record		- 1	- 1

Sex

The sex code was compared between File A and File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	males	22	- 10	12
Partial 2	females	22	- 12	10
Partial 3	female vs male	11	- 70	- 59
Partial 4	both blank	44	- 44	0
Partial 5	male vs blanks	43	- 45	- 2
Partial 6	female vs blanks	43	- 47	- 4

Initials of the First and Second Given Names

All combinations of the initials are compared e.g. first initial together, second initials together and first with second initial.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
init 11	G1=G1,	0		0
init 22	G2=G2	0	- 1	- 1
init 12	G1=G2	0	25	25
init 21	G2=G1	0	25	25

First Given Name

Assign level of agreement for the remaining characters of the first given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	65	- 1	64
Partial 2	4 agree, 3 missing	63	- 1	62
Partial 3	4 agree, 3 disagree	63	- 38	25
Partial 4	3 agree, 4 missing	61	- 1	60
Partial 5	3 agree, 4 disagree	61	- 51	10
Partial 6	2 agree, 5 missing	52	- 1	51
Partial 7	2 agree, 5 disagree	52	- 52	0
Partial 8	1 agree, 6 disagree	38	- 53	- 15
Partial 9	1 agree, 6 missing	38	- 1	37

Second Given Name
Assign level of agreement for the remaining characters of the second given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	65	- 1	64
Partial 2	4 agree, 3 missing	63	- 1	62
Partial 3	4 agree, 3 disagree	63	- 38	25
Partial 4	3 agree, 4 missing	61	- 1	60
Partial 5	3 agree, 4 disagree	61	- 51	10
Partial 6	2 agree, 5 missing	52	- 1	51
Partial 7	2 agree, 5 disagree	52	- 52	0
Partial 8	1 agree, 6 disagree	38	- 53	-15
Partial 9	1 agree, 6 missing	38	- 1	37

First Given Name to Second Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the first given name from File A to the second given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	65	- 6	59
Partial 2	4 agree, 3 missing	63	- 6	57
Partial 3	4 agree, 3 disagree	63	- 38	25
Partial 4	3 agree, 4 missing	61	- 6	55
Partial 5	3 agree, 4 disagree	61	- 51	10
Partial 6	2 agree, 5 missing	52	- 6	46
Partial 7	2 agree, 5 disagree	52	- 52	0
Partial 8	1 agree, 6 disagree	38	- 53	- 15
Partial 9	1 agree, 6 missing	38	- 6	32

Second Given Name to First Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the second given name from File A to the first given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	65	- 8	57
Partial 2	4 agree, 3 missing	63	- 8	55
Partial 3	4 agree, 3 disagree	63	- 38	25
Partial 4	3 agree, 4 missing	61	- 8	53
Partial 5	3 agree, 4 disagree	61	- 51	10
Partial 6	2 agree, 5 missing	52	- 8	44
Partial 7	2 agree, 5 disagree	52	- 52	0
Partial 8	1 agree, 6 disagree	38	- 53	-15
Partial 9	1 agree, 6 missing	38	- 8	30

Surname
Assign level of agreement for the NYSIIS code and surname. Thirteen characters are compared, six for NYSIIS, seven for the surname. Partial agreements occur when not all thirteen characters agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	13, NYSIIS & Surname	80	- 10	70
Partial 2	10, NYSIIS & 4ch. Surname	40	- 20	20
Partial 3	6, NYSIIS agree	20	- 30	-10

Weight Class

Specifies a range of weights which depend on how common or rare the surname is. Higher values are given to rare names. The rules for surname and weight class are used together with the surname providing the negative component and the weight class the positive component.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	Very Common	21		21
Partial 2	Common	32		32
Partial 3	Relatively rare	43		43
Partial 4	Rare	54		54
Partial 5	Very rare	65		65

APPENDIX E

COMPARE RULES FOR NYSIIS LINKAGE

Birth Year
The year of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	55	- 2	53
Partial 2	differ by 1 yr	42	- 45	- 3
Partial 3	differ by 2-3 yrs	32	- 55	- 23
Partial 4	differ by 4-6 yrs	27	- 60	- 33
Partial 5	differ by 7-9 yrs	30	- 70	- 40
Partial 6	differ by 10 yrs	45	- 70	- 25
Disagree			- 60	- 60

Birth Month
The month of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	36	- 3	33
Partial 2	differ by 1 month	27	- 55	- 28
Partial 3	differ by 2-3 months	20	- 55	- 35
Partial 4	differ by 4-6 months	20	- 40	- 20
Partial 5	differ by 11 months	22	- 45	- 23
Disagree	all other cases		- 22	- 22
Missing	blank, either record		- 5	- 5

Birth Day
The day of birth was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	50	- 3	47
Partial 2	differ by 1 day	40	- 50	- 10
Partial 3	differ by 3 days	30	- 45	- 15
Partial 4	differ by 9 days	25	- 60	- 35
Partial 5	differ by 30 days	10	- 60	- 50
Disagree	all other cases		- 50	- 50
Missing	blank, either record		- 5	- 5

Sex
The sex code was compared between records.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	males	23	- 8	15
Partial 2	females	20	- 13	7
Partial 3	female vs male	11	- 65	- 54
Partial 4	both blank	70	- 60	10
Partial 5	male vs blanks	37	- 50	- 13
Partial 6	female vs blanks	37	- 55	- 18

Initials of the First and Second Given Names
All combinations of the initials are compared e.g. first initial together, second initials together and first with second initials.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
	4 initials			
Full	2 agree G1=G1, G2=G2	85	- 19	66
Partial 2	2 agree G1=G2, G2=G1	90	- 60	30
Partial 3	G1=G1 or G2=G2	40	- 55	- 15
Partial 4	G1=G2 or G2=G1	45	- 80	- 35
Partial 5	all disagree	15	- 80	- 75
	3 Initials			
Partial 6	G1=G1	50	- 25	25
Partial 7	G1=G2 or G2=G1	50	- 55	- 5
Partial 8	all disagree	15	- 75	- 60
	2 Initials			
Partial 9	agree	60	- 35	25
Partial 10	disagree	35	- 80	- 45
Missing	blank, either record		- 15	- 15

First Given Name

Assign level of agreement for the remaining characters of the first given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Second Given Name

Assign level of agreement for the remaining characters of the second given names, if initials agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

First Given Name to Second Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the first given name from File A to the second given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	32	- 5	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 64	- 5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Second Given Name to First Given Name (Cross Compare)
Assign level of agreement for the remaining characters of the second given name from File A to the first given name from File B.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	7 characters agree	29	- 2	27
Partial 2	4 agree, 3 missing	81	- 51	30
Partial 3	4 agree, 3 disagree	59	- 54	5
Partial 4	3 agree, 4 missing	87	- 70	17
Partial 5	3 agree, 4 disagree	54	- 67	- 13
Partial 6	2 agree, 5 missing	87	- 70	17
Partial 7	2 agree, 5 disagree	23	- 68	- 45
Partial 8	1 agree, 6 disagree	7	- 54	- 47
Partial 9	1 agree, 6 missing	7	- 4	3

Surname

Assign level of agreement for the NYSIIS code and surname. Thirteen characters are compared, six for NYSIIS, seven for the surname. Partial agreements occur when not all thirteen characters agree.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	13, NYSIIS & Surname	0	- 1	- 1
Partial 2	10, NYSIIS & 4ch. Surname	0	- 40	- 40
Partial 3	6, NYSIIS agree	0	- 60	- 60
Missing	blank, either record			

Weight Class

Specifies a range of weights which depend on how common or rare the surname is. Higher values are given to rare names. The rules for surname and weight class are used together with the surname providing the negative component and the weight class the positive component.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	Very Common	70	- 10	60
Partial 2	Common	70	- 18	52
Partial 3	Relatively rare	90	- 19	71
Partial 4	Rare	105	- 30	75
Partial 5	Very rare	130	- 30	100
Disagree	all other cases		- 25	- 25
Missing	blank, either record	75		75

Inversion of Birth Day with Birth Month

This is a check to see if the value in the birth day field on one file agrees with the value in the birth month field on the other file.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	A=B, and B=A	40	80	120

Birth Place
The birthplace codes are compared. The codes can be of a province or a country level. Some records just have a code denoting Canada.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	30	- 2	28
Partial	Province & Canada	30	- 20	10

Year Last Known Alive
The year last known alive on each record was compared.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	27	- 37	-10

Job Class
The job class on each record is compared.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	35	- 8	27
Disagree	all other cases		- 10	- 10

Year of Entry to Registry
The year of entry to the registry is compared and given a low weight for agreement.

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	40	- 16	24
Disagree	all other cases		- 10	- 10

Last Known Province
The province where the person was last known to work was used for comparison.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	20	- 5	15
Disagree	all other cases		- 20	- 20

Employer Class
The type of employer of the person was used for comparison.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	25	- 5	20
Disagree	all other cases		- 15	- 15

S.I.N.

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	6	- 70	- 64
Partial 2	both blank	40	- 15	25
Partial 3	one blank	15	- 12	3
Disagree	all other cases		- 35	- 35

RPB

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	- 20		- 20

Namecode

Level of Agreement	Condition	Weights Positive	Negative	Net
Full	equal	- 20		- 20

Combseq

Level of Agreement	Condition	Weights		
		Positive	Negative	Net
Full	equal	- 20		- 20